

Estadística Descriptiva

Estadística — ITT Sonido e Imagen — curso 2004-2005

1. Definiciones fundamentales

La *Estadística Descriptiva* se ocupa de la descripción de datos experimentales, más específicamente de la recopilación, organización y análisis de datos sobre alguna característica de ciertos individuos pertenecientes a la *población* o *universo*.

Definición 1 (Población, tamaño). Llamamos **población** a un conjunto bien definido sobre el que se observa o puede observarse una cierta característica. Puede ser finita o infinita. El **tamaño de la población** es el número de individuos que tiene, su cardinal, lo denotamos por N .

Si la población es muy grande se hace muy costoso y en algunos casos imposible considerar cada *individuo* y se realiza una selección denominada *muestra*.

Definición 2 (Individuo). Llamamos **individuo** a cada uno de los elementos de la población.

Definición 3 (Muestra, tamaño). Una **muestra** es un conjunto de individuos de la población que refleja las características de ésta lo mejor posible. Si las características quedan bien reflejadas, se dice que la muestra es *representativa*. El **tamaño de una muestra** es el número de individuos que tiene, lo denotamos por n .

Si muestra y población coinciden, se dice que se dispone de un **censo**.

Definición 4 (Variable, dato). Una **variable** (X) es un símbolo que representa una característica a estudiar en la población. Llamamos **dato** (x) al valor (numérico o no) que la variable toma sobre un individuo concreto de la muestra.

Tipos de variables

- **Cuantitativa:** toma valores en un conjunto prefijado de valores numéricos, se puede medir.
 - **Discreta:** el conjunto es finito o numerable (Ej. número de hijos de una familia).
 - **Continua:** el conjunto es infinito no numerable, contiene algún intervalo (Ej. duración de alguna componente en un sistema).
- **Cualitativa:** toma valores que se corresponden con cualidades no cuantificables de los individuos, no se pueden medir (Ej. color).
- **Dicotómicas:** sólo pueden tomar dos valores, (SI/NO); (0,1).

2. Representaciones tabulares, frecuencias

Una vez obtenida una muestra de cualquier población y observados los valores que toma la variable en los individuos de la muestra, estos valores se suelen ordenar. Si la variable es cuantitativa la ordenación será de menor a mayor.

Dada una variable X , consideramos una muestra de tamaño n que toma k valores distintos, x_1, \dots, x_k (si la variable es cuantitativa $x_1 < x_2 < \dots < x_k$).

La **frecuencia absoluta de un valor** x_i es el número de veces que dicho valor aparece en la muestra. Se representa por n_i y cumple

$$\sum_{i=1}^k n_i = n_1 + n_2 + \dots + n_k = n$$

La **frecuencia relativa de un valor** x_i es el cociente de la frecuencia absoluta de x_i (n_i) entre el tamaño de la muestra (n), se representa por f_i

$$f_i = \frac{n_i}{n}, \quad \text{se cumple} \quad \sum_{i=1}^k f_i = 1.$$

Si trabajamos con variables cuantitativas, como hemos ordenado los valores de la muestra de menor a mayor, podemos definir las frecuencias acumuladas.

La **frecuencia absoluta acumulada del valor i -ésimo** es la suma de las frecuencias absolutas hasta dicho valor, se denota por N_i

$$N_i = n_1 + n_2 + \dots + n_i$$

La **frecuencia relativa acumulada del valor i -ésimo** es la suma de las frecuencias relativas hasta dicho valor, se denota por F_i

$$F_i = f_1 + f_2 + \dots + f_i, \quad F_i = \frac{N_i}{n}$$

Una **tabla de frecuencias** tiene la siguiente estructura.

x_i	n_i	f_i	N_i	F_i

Podríamos hablar también de la frecuencia de un cierto valor dentro de una población (siempre que ésta fuera finita), bastaría con tomar como muestra un censo. Lo mismo ocurre para todas las medidas que describiremos más adelante (de tendencia central, posición, dispersión y forma), en principio nos referiremos a medidas sobre una muestra, en otro caso (si fueran relativas a la población) lo explicitaríamos.

3. Datos agrupados

A veces se hace necesario trabajar con datos agrupados (el por qué y cómo fueron brevemente explicados en clase). Definimos entonces como **clase** a cada uno de los intervalos en que se agrupan los datos. Las frecuencias harán ahora referencia al número de datos que hay en cada intervalo.

Denotaremos la i -ésima clase como $[\underline{L}_i, \overline{L}_i]$. Si sucede que $\underline{L}_i = \overline{L}_{i-1}$, las clases serán de la forma $(\underline{L}_i, \overline{L}_i]$, de tal modo que la intersección de dos clases distintas sea el vacío.

Dada la i -ésima clase, \underline{L}_i será su **límite inferior** y \overline{L}_i su **límite superior**. La **marca de clase** será el punto medio del intervalo, $m_i = (\underline{L}_i + \overline{L}_i)/2$ y la **amplitud** el tamaño del intervalo, $c_i = \overline{L}_i - \underline{L}_i$.

Una **tabla de frecuencias** tendrá ahora la siguiente estructura.

$(\underline{L}_i, \overline{L}_i]$	n_i	f_i	N_i	F_i

4. Medidas de posición

4.1. Medidas de tendencia central

Los promedios o medidas de tendencia central son valores típicos o representativos de un conjunto de datos. Pretenden resumir todos los datos en un único valor. Definimos tres medidas de tendencia central, media, mediana y moda.

4.1.1. Media, (\bar{x})

Se calcula para variables cuantitativas y se trata del centro geométrico o de gravedad de nuestros datos,

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \dots + n_kx_k}{n} = \sum_{i=1}^k x_i f_i$$

Si se trata de una media poblacional, es decir, estamos considerando todos los individuos de la población, suele denotarse por μ .

Propiedades.

1. $\sum_{i=1}^k (x_i - \bar{x})n_i = 0$
2. la media es el punto para el que la distancia cuadrática media a los valores de la muestra es mínima, es decir, para cualquier $a \in \mathbb{R}$

$$\sum_{i=1}^k (x_i - \bar{x})^2 n_i \leq \sum_{i=1}^k (x_i - a)^2 n_i$$

Obsérvese que en el cómputo de la media se utilizan todos los valores, por tanto si hay valores anómalos (extremos) influirán fuertemente en ella.

Si los datos están agrupados, para hallar la media tomamos la marca de las clases,

$$\bar{x} = \frac{n_1m_1 + n_2m_2 + \dots + n_km_k}{n} = \sum_{i=1}^k m_i f_i$$

4.1.2. Mediana, (Me)

Se calcula para variables cuantitativas, es un número tal que al menos el 50 % de los datos es menor o igual que la mediana y al menos el 50 % mayor o igual. Si hay más de una mediana tomamos el punto medio entre la mediana mayor y la más pequeña, que serán los datos que aparecen en la muestra y sirven como medianas.

Para calcularla, recurrimos a la columna de las frecuencias relativas acumuladas y buscamos el primer valor $F_i \geq 0.5$, es decir aquel para el que $F_i \geq 0.5$ y $F_{i-1} < 0.5$. Si $F_i > 0.5$, entonces $Me = x_i$, si $F_i = 0.5$, entonces $Me = (x_i + x_{i+1})/2$.

Propiedad. La mediana es el punto para el que la distancia euclídea media a los valores de la muestra es mínima, es decir, para cualquier $a \in \mathbb{R}$

$$\sum_{i=1}^k |x_i - Me| n_i \leq \sum_{i=1}^k |x_i - a| n_i$$

Sólo tiene en cuenta la posición de los valores en la muestra y por lo tanto tiene mucho mejor comportamiento que la media cuando hay observaciones anómalas.

4.1.3. Moda, (Moda)

Es el valor con mayor frecuencia. Si hay más de una, la variable se dice multimodal y puede calcularse para cualquier tipo de variable.

Si los datos están agrupados hablamos de clase modal y será aquella para la que el cociente frecuencia relativa dividido entre amplitud (f_i/c_i) es mayor.

4.1.4. Media armónica, (\bar{x}_H)

$$\bar{x}_H = \frac{n}{\sum_{i=1}^k n_i / x_i}$$

4.1.5. Media geométrica, (\bar{x}_G)

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$$

4.1.6. Media recortada al 5 %, (\bar{x}_R)

$$\bar{x}_R = \frac{1}{0.9} \left[(F_{k_1} - 0.05)x_{k_1} + (0.95 - F_{k_2-1})x_{k_2} + \sum_{i=k_1+1}^{k_2-1} f_i x_i \right]$$

con k_1 y k_2 satisfaciendo,

$$F_{k_1-1} < 0.05 \leq F_{k_1} \quad ; \quad F_{k_2-1} \leq 0.95 < F_{k_2}$$

4.2. Cuantiles

Se calculan para variables cuantitativas y al igual que la mediana sólo tienen en cuenta la posición de los valores en la muestra. Casos particulares de cuantiles son los *cuartiles*, los *percentiles* y los *deciles* (estos últimos dividen la muestra ordenada en 10 partes).

4.2.1. Cuartiles

Dividen la muestra ordenada en 4 partes.

- Q_1 , primer cuartil, al menos el 25 % de los datos son menores o iguales que él y al menos el 75 % de los datos son mayores o iguales que él.
- Q_2 , segundo cuartil, es la mediana, $Q_2 = \text{Me}$.
- Q_3 , tercer cuartil, al menos el 75 % de los datos son menores o iguales que él y al menos el 25 % de los datos son mayores o iguales que él.
- Q_4 , cuarto cuartil, es el mayor valor que se alcanza en la muestra.

4.2.2. Percentiles

Dividen la muestra ordenada en 100 partes.

Dado $\alpha \in \mathbb{N}$ tal que $1 \leq \alpha \leq 99$, el α -ésimo percentil, P_α es un valor tal que al menos el α % de los datos son menores o iguales que él y al menos el $(100 - \alpha)$ % de los datos son mayores o iguales que él.

A partir de las definiciones de los cuartiles y percentiles, es claro que $Q_1 = P_{25}$ y $Q_3 = P_{75}$.

Para calcular el percentil P_α , buscamos en la columna de las frecuencias relativas acumuladas el primer valor mayor o igual que $\alpha/100$, es decir, buscamos $F_i \geq \alpha/100$ tal que $F_{i-1} < \alpha/100$. Si $F_i > \alpha/100$, entonces $P_\alpha = x_i$, si $F_i = \alpha/100$, entonces $P_\alpha = (\alpha/100)x_i + (1 - \alpha/100)x_{i+1}$.

5. Medidas de dispersión

Sólo tienen sentido para variables cuantitativas y las definimos para variables no agrupadas.

5.1. Recorrido o rango

Diferencia entre el mayor y menor valor de una muestra, $x_k - x_1$.

5.2. Rango semiintercuartílico y amplitud intercuartil

El rango semiintercuartílico es la mitad de la diferencia entre el tercer y primer cuartil, $Q = (Q_3 - Q_1)/2$. La amplitud intercuartil es el doble del valor anterior, $2Q = (Q_3 - Q_1)$.

5.3. Desviación típica, (s)

Cuantifica el error que cometemos si representamos una muestra únicamente por su media.

$$s = \sqrt{\sum_{i=1}^k (x_i - \bar{x})^2 f_i}$$

La desviación típica poblacional suele denotarse por σ .

5.4. Varianza muestral, (s^2)

$$s^2 = \sum_{i=1}^k (x_i - \bar{x})^2 f_i$$

La varianza poblacional suele denotarse por σ^2 .

Propiedad.

$$s^2 = \sum_{i=1}^k x_i^2 f_i - (\bar{x})^2$$

5.5. Cuasivarianza muestral, (s^{*2})

$$s^{*2} = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{n - 1} = \frac{n}{n - 1} s^2$$

5.6. Desviación media respecto de la mediana

$$DM = \frac{\sum_{i=1}^k |x_i - Me| n_i}{n}$$

5.7. Coeficiente de variación, (CV)

$$CV = \frac{s}{|\bar{x}|} 100$$

Las medidas de dispersión anteriores dependen de las unidades de medida, el coeficiente de variación es, en cambio, una medida de dispersión relativa (adimensional).

También existen las llamadas **medidas de forma** que nos indican numéricamente cómo están distribuidos los datos en una muestra.

6. Medidas de forma

6.1. Asimetría

El coeficiente de asimetría de una variable mide el grado de asimetría de la distribución de sus datos en torno a su media. Es adimensional y se define como sigue:

$$As = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^3 / n}{s^3}.$$

Las colas de una variable están constituidas por los valores alejados de la media (valores extremos). Una variable es asimétrica si su cola a un lado más larga que su cola al otro y simétrica si ambas colas son igual de largas.

- si $As > 0$ la distribución será asimétrica a la derecha. La cola a la derecha es más larga que la cola a la izquierda.
- si $As = 0$ la distribución será simétrica. Ambas colas son igual de largas
- si $As < 0$ la distribución será asimétrica a la izquierda. La cola a la izquierda es más larga que la cola a la derecha.

6.2. Apuntamiento o curtosis

El coeficiente de apuntamiento o curtosis de una variable sirve para medir el grado de concentración de los valores que toma en torno a su media. Se elige como referencia una variable con distribución normal, de tal modo que para ella el coeficiente de apuntamiento es 0.

$$Ap = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^4 / n}{s^4} - 3.$$

Según su apuntamiento, una variable puede ser:

- **Leptocúrtica**, si $Ap > 0$, es decir, es más apuntada que la normal. Los valores que toma la variable están muy concentrados en torno a su media y hay pocos valores extremos.
- **Mesocúrtica**, si $Ap = 0$, es decir, es tan apuntada como la normal.
- **Platicúrtica**, si $Ap < 0$, es decir, es menos apuntada que la normal. Hay muchos valores extremos, las colas de la variable son muy pesadas.

7. Representaciones gráficas

7.1. Diagrama de barras

En el eje OX representamos los valores de las variables y levantamos un trazo o barra de longitud igual a la frecuencia relativa (o absoluta).

7.2. Pictogramas

Figuras cuya área es la frecuencia (o un valor proporcional) del valor que representan.

7.3. Diagrama de sectores

Se divide un círculo en sectores cada uno de ellos proporcional a la frecuencia relativa de un valor.

7.4. Histograma

Es la representación más frecuente con datos agrupados. Está formado por un conjunto de rectángulos tales que:

1. Sus bases coinciden con el intervalo que representan y cuyos valores aparecen en el eje OX .
2. El área de cada rectángulo debe ser igual a la frecuencia relativa del intervalo. Su altura será por tanto f_i/c_i y la suma de las áreas de todos los rectángulos la unidad.

7.5. Polígono de frecuencias (poligonal de frecuencias)

Se obtiene uniendo los puntos medios de los extremos superiores de los rectángulos que forman el histograma, es decir los puntos $(m_i, f_i/c_i)$. En los extremos, unimos $(m_1, f_1/c_1)$ con $(\underline{L}_1, 0)$ y $(m_k, f_k/c_k)$ con $(\overline{L}_k, 0)$.

7.6. Diagrama de tallos y hojas

Procedimiento semigráfico para el que se preparan los datos resumiéndolos en dos o tres cifras (expresándolos en las unidades adecuadas). A continuación se disponen en una tabla de dos columnas del siguiente modo:

1. Si los datos son de dos dígitos, a la izquierda (en el tallo) aparece la cifra de las decenas, a la derecha separadas por una línea aparecen las hojas y se escriben todas seguidas.
2. Si hay tres dígitos el tallo está formado por los dos primeros.

Ejemplo. Dada la muestra $\{114, 125, 114, 124, 152, 134\}$, dibuja su diagrama

de tallos y hojas.

10		
11		4 4
12		4 5
13		4
14		
15		2

, las hojas son las unidades

Observación. Se trata de un histograma con amplitud de las clases constante y girado 90^0 .

7.7. Diagrama de cajas

Paralelo a un eje numerado dibujamos un segmento con extremos en los valores menor y mayor que aparecen en la muestra y que marcamos con dos *bigotes*. Dibujamos además una caja con extremos en el primer y tercer cuartil y marcamos en ella la mediana.

Observación. En los diagramas de cajas que nos ofrecen ciertos paquetes estadísticos aparecen reflejados los *valores atípicos* y *casos extremos* fuera del segmento.

8. Estadística descriptiva bidimensional

Estudiamos simultáneamente dos variables del individuo.

Definición 5. Una **variable bidimensional** (X, Y) es un símbolo que representa dos características de los individuos de la población.

Dada una variable bidimensional (X, Y) , consideramos una muestra de tamaño n en la que X toma k valores distintos, x_1, \dots, x_k , si la variable es cuantitativa $x_1 < x_2 < \dots < x_k$ e Y toma l valores distintos, y_1, \dots, y_l , si la variable es cuantitativa $y_1 < y_2 < \dots < y_l$.

Obtenemos, por tanto, observaciones del tipo (x_i, y_j) .

La **frecuencia absoluta de un valor** (x_i, y_j) es el número de veces que dicho valor aparece en la muestra. Se representa por n_{ij} , se cumple

$$\sum_{i=1}^k \sum_{j=1}^l n_{ij} = n.$$

La **frecuencia relativa de un valor** (x_i, y_j) es el cociente de la frecuencia absoluta de (x_i, y_j) , n_{ij} entre el tamaño de la muestra n , se representa por f_{ij}

$$f_{ij} = \frac{n_{ij}}{n} \quad \text{se cumple} \quad \sum_{i=1}^k \sum_{j=1}^l f_{ij} = 1.$$

8.1. Distribuciones marginales

Nos indican el comportamiento aislado de cada una de las variables X e Y que dan lugar a una variable bidimensional.

Frecuencia absoluta marginal de x_i , $n_{i.} = n_{i1} + n_{i2} + \dots + n_{il} = \sum_{j=1}^l n_{ij}$.

Frecuencia relativa marginal de x_i , $f_{i.} = n_{i.}/n$.

Frecuencia absoluta marginal de y_j , $n_{.j} = n_{1j} + n_{2j} + \dots + n_{kj} = \sum_{i=1}^k n_{ij}$.

Frecuencia relativa marginal de y_j , $f_{.j} = n_{.j}/n$.

Podemos calcular las medidas de tendencia central o dispersión y realizar cualquier tipo de representación gráfica de las marginales.

Una **tabla de doble entrada** de una variable bidimensional sigue la estructura que se presenta a continuación, en la que tienen cabida las frecuencias marginales (representadas en la última fila y última columna). Puede ser de frecuencias absolutas o relativas.

$X \backslash Y$	y_1	y_2	\dots	y_l	$n_{i.}$
x_1	n_{11}	n_{12}	\dots	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	\dots	n_{2l}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_k	n_{k1}	n_{k2}	\dots	n_{kl}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	\dots	$n_{.l}$	n

8.2. Distribuciones condicionadas

Son distribuciones fijada una condición. Dicha condición puede ser sobre la misma variable o la otra.

La **frecuencia absoluta de x_i dada cierta condición** será el número de observaciones en la muestra que cumplen la condición y para las que la variable X toma el valor x_i .

La **frecuencia relativa de x_i dada cierta condición** será la frecuencia absoluta de x_i dada la condición dividida entre el número total de observaciones de la muestra que cumplen la condición.

Podemos hablar de la distribución de X condicionada a que Y toma el valor y_j , $X|_{Y=y_j}$ y será la distribución de todas las observaciones con valor y_j en Y . Su distribución de frecuencias absolutas ($n_{i|j}$) será la columna j -ésima de la tabla de doble entrada, las frecuencias relativas vendrán dadas por $f_{i|j} = n_{ij}/n_{.j}$.

Podemos hablar de medidas de tendencia central o dispersión para distribuciones marginales.

8.3. Independencia estadística

El interés del estudio conjunto de dos variables como variable aleatoria bidimensional es sacar conclusiones sobre la posible relación de dependencia entre ellas. Dos variables son estadísticamente independientes cuando no existe relación alguna entre ellas.

Definición 6. Dos variables X e Y se dicen **independientes** si las distribuciones de X condicionadas a cualquier valor de Y son iguales, es decir,

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \dots = \frac{n_{il}}{n_{.l}} \quad \text{para todo } i = 1, \dots, k$$

o equivalentemente

$$f_{i|1} = f_{i|2} = \dots = f_{i|l} \quad \text{para todo } i = 1, \dots, k$$

Se puede demostrar que la relación anterior es equivalente a

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n} \quad \text{para todo } i, j.$$

Es decir, las variables X e Y son estadísticamente independientes si la frecuencia relativa conjunta de cada par de valores es igual al producto de las frecuencias relativas marginales ($f_{ij} = f_{i.}f_{.j}$ para todo i, j).

Comentario. El valor esperado de la casilla (i, j) si las variables fuesen independientes se obtiene utilizando la fórmula $nf_{i.}f_{.j}$

8.4. Regresión lineal (método de mínimos cuadrados), correlación

En este apartado consideraremos que las variables con las que trabajamos son cuantitativas.

8.4.1. Nube de puntos o diagrama de dispersión

El procedimiento gráfico habitual para representar una variable bidimensional es una **nube de puntos** o **diagrama de dispersión** en la que cada valor (x_i, y_j) que aparece en la muestra se representa por un único punto de abscisa x_i y ordenada y_j . En dicha nube de puntos podemos apreciar la relación entre las variables.

8.4.2. Covarianza, (s_{XY})

Definimos la **covarianza** de una variable bidimensional (X, Y) como:

$$s_{XY} = \frac{\sum_{i=1}^k \sum_{j=1}^l (x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}.$$

Propiedad.

$$s_{XY} = \frac{\sum_{i=1}^k \sum_{j=1}^l x_i y_j n_{ij}}{n} - \bar{x} \bar{y}.$$

- Si la covarianza es positiva ($s_{XY} > 0$), existirá tendencia a que las mayores observaciones de una de las variables se correspondan con las mayores observaciones de la otra.
- Si la covarianza es negativa ($s_{XY} < 0$), existirá tendencia a que las mayores observaciones de una de las variables se correspondan con las menores de la otra.
- Si la covarianza es cero ($s_{XY} = 0$), no existe relación **lineal** entre las variables.

Si X e Y son independientes, entonces su covarianza será cero, $s_{XY} = 0$, el resultado recíproco no es cierto.

8.4.3. Regresión lineal, mínimos cuadrados

La **regresión** consiste en modelizar la relación de dependencia entre las variables y predecir los valores de una de ellas (**variable dependiente**) en función de los valores de la otra (**variable independiente o explicativa**).

La manera de conseguir este objetivo es ajustar una ecuación dada a la nube de puntos, en nuestro caso como la regresión es lineal, la ecuación será la de una recta.

Sea X la variable independiente e Y la variable dependiente, planteamos la ecuación de una recta $\hat{y} = a + bx$ para estimar Y a partir de X . Buscamos los valores a, b para los que la suma del error cuadrático es más pequeña, es decir, dada la función

$$F(a, b) = \sum_{i=1}^k \sum_{j=1}^l (y_j - (a + bx_i))^2 n_{ij}$$

queremos hallar los valores de a y b para los que $F(a, b)$ es más pequeña.

Dichos valores son

$$b = \frac{s_{XY}}{s_X^2}, \quad a = \bar{y} - \frac{s_{XY}}{s_X^2} \bar{x}.$$

A b le llamamos **coeficiente de regresión lineal de Y sobre X** . Con a y b obtenemos la **recta de regresión de Y sobre X** , que expresada en su ecuación punto-pendiente resulta ser

$$\hat{y} - \bar{y} = \frac{s_{XY}}{s_X^2} (x - \bar{x}),$$

es decir, la pendiente de la recta de regresión de Y sobre X es el coeficiente de regresión lineal de Y sobre X y pasa por el punto que tiene por abscisa la media de X y por ordenada la media de Y , (\bar{x}, \bar{y}) .

La recta de regresión de X sobre Y se calcula de modo análogo a la de Y sobre X .

8.4.4. Correlación lineal

El **coeficiente de correlación lineal de Pearson** se define como

$$r = \rho_{XY} = \frac{s_{XY}}{s_X s_Y}$$

y toma valores entre -1 y 1 .

- Si $r = 1$ decimos que hay correlación positiva perfecta.
- Si $r = -1$ decimos que hay correlación negativa perfecta.
- Si $r < 0$ hay correlación lineal negativa, ambas rectas de regresión son decrecientes.
- Si $r > 0$ hay correlación lineal positiva, ambas rectas de regresión son crecientes.
- Si $r = 0$ las variables son incorreladas o linealmente independientes.

El **coeficiente de determinación lineal** o de **bondad de ajuste** es el cuadrado del coeficiente de correlación, r^2 , está en el intervalo $[0, 1]$ y cuanto mayor sea, mejor será el ajuste.